EUROPEAN JOURNAL OF STEM EDUCATION

European Journal of STEM Education,
2024, 9(1), 06
ISSN: 2468-4368

lectito

Review Article

A Review of the Effects of Integrated Language, Science and Technology Interventions in Elementary Education on Student Achievement

Miriam J. Rhodes [1]*, Adrie J. Visscher [1], Hanno van Keulen [2], Martine A. R. Gijsel [3]

[1] *University of Twente*, THE NETHERLANDS
[2] *Delft University of Technology*, THE NETHERLANDS
[3] *Saxion University of Applied Sciences*, THE NETHERLANDS

***Corresponding Author:** m.j.rhodes@utwente.nl

Citation: Rhodes, M. J., Visscher, A. J., van Keulen, H. and Gijsel, M. A. R. (2024). A Review of the Effects of Integrated Language, Science and Technology Interventions in Elementary Education on Student Achievement. *European Journal of STEM Education*, 9(1), 06. https://doi.org/10.20897/ejsteme/14570

Published: May 4, 2024

ABSTRACT

This systematic literature review presents a review of the effects of integrated language arts, science and technology (ILS&T) instruction, with an inquiry- or design-based pedagogy, in elementary schools on student achievement. To this end, an overview of the characteristics of the 19 included studies and their interventions is first presented. Second, the effects of interventions in relation to the study characteristics and outcome variables were examined, by comparing the mean effect sizes. The findings demonstrate positive effects of ILS&T instruction for all reported student learning outcome variables. Third, the relation between characteristics of the intervention and effect sizes was analysed. Interventions with higher levels of integration, a short duration, and teacher professional development activities produce higher effect sizes. These findings are relevant for the design of ILS&T interventions. The analysis was challenged by a lack of detailed information in study and intervention descriptions, which prompts a call for scholars to provide more comprehensive information in their intervention studies.

Keywords: science and technology instruction, inquiry and design based education, language arts, integrated curricula, elementary education, systematic literature review

INTRODUCTION

In the last decades, an increasing number of interventions describe the integration of language arts and inquiry-based science and technology (S&T) instruction in elementary education (Cervetti et al., 2012; Vitale and Romance, 2011). In S&T education, students are taught about the natural and material (human-made) environment and technological artifacts by collecting, analysing and interpreting information gathered through experimenting and testing (Cakir, 2008; NGSS Lead States, 2013). Although the precise elementary school language arts curriculum varies across schools and countries, curriculum core standards typically address components of reading skills, writing skills and oral language in the first language (e.g., see Common Core Standards Initiative, 2010; International Reading Association and National Council of Teachers of English, 1996; SLO, 2020). Integrating language arts and inquiry-based S&T education aligns with sociocultural theories of learning, which emphasize that language learning can be enhanced when it occurs in a socially and culturally relevant context to students (Lemke, 1990). In line with these theories, science education is viewed as a community of discourse (Leach and Scott, 1995;

Lemke, 1990), as language plays a vital role in developing an understanding of the world, and more specifically, of S&T (Daniels, 2001; Lemke, 1990). Constructivist and sociocultural theories of learning emphasize the importance of inquiry- or design-based pedagogies in S&T education, as it suits the true nature of S&T (Lewis, 2006). The theoretical alignments between S&T and language learning are reflected in recent curricular shifts towards promoting science knowledge-in-use (Harris et al., 2019) and language-in-use (Lee et al., 2013), rather than focusing on the body of knowledge of the subjects. Recent studies have showed there are multiple connections between the standards of the two subjects (Lee et al., 2013). Language arts and S&T instruction involve many similar cognitive and intellectual processes, such as making predictions and assessing the quality of arguments and assumptions based on data and evidence (Baker, 1991; Bradbury, 2014). Consequently, interest in linking the curricular contents of S&T with language arts education has increased in recent years.

For teachers, an important motive for integrating instruction in language arts and S&T concerns the limited time spent on teaching S&T in elementary education (Martin et al., 2012). Many teachers feel insecure about their content knowledge and scientific and technological skills (Appleton, 2007; Asma et al., 2011; Traianou, 2007). Integrating instruction in S&T and language arts can make S&T instruction more appealing for teachers, and could thus increase the time spent on teaching S&T (Appleton, 2007).

In response to these developments, interventions have been designed to evaluate the effects of integrated language arts and S&T (ILS&T) curricula that adopt an inquiry- or design-based pedagogy (e.g., Guthrie et al., 2004; Romance and Vitale, 2001). Because of the natural connection between both subjects and the potential learning opportunities in integrated approaches, an integrated approach could promote students' metacognitive and conceptual growth, and lead to a higher level of retention (Yore and Treagust, 2006). However, interventions have shown high levels of variation with respect to their content and instructional approach, and in terms of their effects on students' learning outcomes. An analysis of the differences between the interventions is needed to understand the variability in their effects. Integrated language arts and S&T curricula may differ substantially, because of the various language arts modalities (i.e., reading, writing, oral language, and vocabulary) and S&T characteristics (e.g., design or inquiry, primary focus on knowledge or skills) that can be involved. The characteristics of the interventions will logically affect their results. Below, we explicate how this review differs from previous reviews on this topic. This review aims to provide a systematic analysis of the effects of the reported ILS&T interventions by answering the following research questions:

1. What features characterize studies of and interventions for ILS&T?
2. Does ILS&T instruction enhance language arts and S&T learning compared to language arts and S&T instruction that is not integrated?
3. Which characteristics of ILS&T interventions are associated with the effects on language arts and S&T learning?

BACKGROUND LITERATURE

Before tackling the empirical literature related to the research questions, it is necessary to consider what the relevant theory and research have to say about expected student learning gains as a result of learning language arts and S&T in an integrated way. Given those expected learning gains, it is also important to develop a theoretical and empirical basis for determining which study and intervention characteristics could moderate the effects on student achievement in language arts and S&T.

Expected Student Learning Outcomes of ILS&T Instruction

From a theoretical perspective, ILS&T instruction might improve learning outcomes with respect to several aspects of language arts and S&T learning.

Children's language skills might benefit from the meaningful and authentic context of science education. Integration of the language arts and science S&T curriculum provides students with an authentic purpose for communicating and using and interpreting language in different forms (e.g., texts, conversations, figures, see Christie, 2017; Hapgood and Palincsar, 2006). Because of this, scholars have proposed that S&T instruction can create a stronger sense of purpose and relevance and can therefore facilitate retention for language learning (Guthrie et al., 2006; Stoller, 2008). Scholars have also argued that reading motivation can be enhanced by situating reading in a meaningful and activating context, such as the science classroom (Wigfield and Guthrie, 1997).

Moreover reading (and writing) can be viewed as a constructive process of interpreting disciplinary information (Osborne, 2002), and prior knowledge is therefore considered to be a critical determinant of disciplinary learning, including in the language arts (Kintsch, 2004).

From the perspective of science education, S&T learning can arguably benefit from integration with language arts instruction, because more advanced language skills can help students transform global ideas into (S&T)

knowledge that is more coherent and structured (Osborne, 2002). S&T knowledge and skills strongly appeal to students' (productive and receptive) language skills. Like in the scientific community, students have to communicate and interpret information in texts and discussions that is often abstract and complex (for instance, see Lee et al., 2013). Enhancing students' language skills can therefore also enhance students' ability to participate in S&T practices.

Finally, various scholars have argued that students' attitudes towards S&T learning could be enhanced by making S&T instruction more relevant and coherent to students' daily lives, and by focusing on the distinctive value of engaging in science (Chen et al., 2014; Jenkins, 2011). This can be realized through integration with language arts education, by offering real-world examples through texts and discussions, and engaging with scientific content in meaningful and multimodal ways.

Although integration can be an effective approach to teaching language arts and S&T, there are also potential obstacles. Nixon and Akerson (2004) argued that it may be difficult to align the complex goals of language arts learning with the goals of S&T. For example, when students are instructed to report about their scientific observations using a new writing framework, this may end up suppressing the cognitive processing of science concepts, because students' attention is mainly directed towards the correct use of the new writing structure. The limited attentional capacity model suggests that increasing the cognitive complexity of a task may cause students to prioritize one aspect of performance and neglect the others, due to limited attentional resources (Skehan and Foster, 2001). Similarly, teachers may feel the need to make allowances for the learning objectives of one of the domains to make the learning activities feasible for the students. This may lead to superficial treatment of challenging learning material, which can be demotivating for students (Brophy and Alleman, 1991).

Prior Reviews and Current Study

Scholars have previously examined the added value of curricula that integrate science and language for student learning. Yore et al. (2003) conducted an extensive conceptual review of the literature on literacy and science integration in order to outline current trends and future directions for this area of research, but did not evaluate the effects on student learning. Bradbury (2014) reviewed the literature on science and language integration, with an emphasis on empirical studies and their effects, but did not distinguish between traditional science instruction and inquiry-based or design-based instruction. Furthermore, the review did not report the ESs of the studies, but rather presented a descriptive analysis of the interventions. Several meta-analyses have evaluated the effect of literacy and science integration on specific aspects of language arts learning, such as vocabulary (Guo et al., 2016) and writing (Graham et al., 2020). Other meta-analyses have examined the integration of language arts instruction with other content-area learning (Graham et al., 2020; Hwang et al., 2022), including science, social studies, and mathematics. Although these reviews reported promising results, it is not yet possible to disentangle the effects of ILS&T instruction based on their findings. In particular, the impact of ILS&T instruction with inquiry- or design-based pedagogy has not been previously subjected to thorough examination in any review. This pedagogical focus is relevant, because of the shift from traditional, teacher-led science instruction to inquiry- or design-based pedagogies in curricular standards (e.g., NGSS Lead States, 2013). These pedagogies are underpinned by different theories of learning (e.g., behaviourism, constructivism) and are characterized by different beliefs about how knowledge is constructed. Furthermore, the way in which learning content is offered to students impacts learning outcomes (for a recent review, see Khalaf and Mohammed Zin, 2018). Inquiry- or design-based learning also exposes students to discipline-specific language and emphasizes communicating about complex concepts and relationships. The current review additionally encompasses all domains of language learning (i.e., reading, writing, oral language, and vocabulary) to reflect the language arts curriculum in elementary schools. In this way, the effects that are described in these interventions are closely aligned with classroom practice advocated in many current educational standards (e.g., NGSS Lead States, 2013). Finally, scholars have not previously canvassed the features of studies and interventions to identify effective approaches to ILS&T instruction. This requires a clear characterization of ILS&T studies and interventions, and an analysis of the relation between these characteristics and the intervention effects, which was a second aim of the current review.

Potential Moderators

Various factors can potentially affect the impact of ILS&T interventions on student outcomes. The effects on student outcomes can vary depending on the characteristics of the study (e.g., study design, instruments used to measure student achievement) and of the intervention (e.g., learning goals, instructional method). Potential moderators are described below and summarized in **Table 2**.

Study characteristics

When comparing effects of educational interventions, it is first important to consider the research design of the study (Wilson and Lipsey, 2001). Studies with a (cluster) randomized design (i.e., experiments in which clusters of people rather than individuals are assigned at random to treatments, such as pre-existing classes, schools) are generally preferred for studying intervention effects, as this design affords the best ground for causal inferences. When experimental and control groups are not assigned at random, this may lead to experimental and control groups that are not on average equal on relevant characteristics (e.g., motivation, performance level) which makes it difficult to attribute results to the intervention. In quasi-experiments, matching procedures can be used to promote that participants in the two conditions are as much as possible similar with respect to certain covariates (e.g., demographics, ability level, school setting). However, matching is only possible for those variables one has information on. It is, for example, difficult to match teachers on their teaching ability (Borenstein et al., 2019).

Second, it is important to consider the type of control group that is included in experimental studies. The effect of ILS&T instruction can be compared to control groups receiving only language instruction, only S&T instruction, or separate language and S&T instruction. For instance, when comparing the ILS&T intervention to only S&T instruction, students in the control group receive the same S&T instruction as the experimental group, but without the incorporation of language instruction. Therefore, it would seem likely that the control group students would show progress on their S&T outcomes, but not so much on their language outcomes, as they did not receive that instruction at all. Thus, the comparison with each type of control group presents a slightly different type of evidence.

Third, the implementation scale can affect the outcome of an intervention. Studies with a smaller sample size make it easier for program developers to maintain a higher degree of treatment fidelity ("super implementation"), and therefore tend to overestimate effects (Cheung and Slavin, 2016).

Fourth, the methods used to evaluate the intervention effects should be considered. Independent instruments are preferred to researcher-developed instruments (Wolf and Harbatkin, 2023). Researcher-developed instruments specifically designed for a study are often associated with higher effect sizes (ESs), because of their (over)alignment with the intervention (Cheung and Slavin, 2016; Wilson and Lipsey, 2001). The time of testing in pre-post-test designs can also influence intervention effects. Post-test measures that are administered directly after the intervention only inform about the short-term effects, whereas effects found with retention tests are more likely to be long-lasting. In addition, when reporting ESs, post-test gains that are not adjusted for pre-test differences can be a lingering effect of pre-test differences or standard error, instead of a treatment effect (Wilson and Lipsey, 2001).

Intervention characteristics

Besides the study characteristics, the interventions described in the studies can also vary, in terms of both the instructional intervention provided to students and the support (if any) given to teachers via a teacher professional development (TPD) program.

Instructional intervention. It is crucial to consider instructional alignment, which has been demonstrated to produce higher achievement results than poorly aligned instruction (Cohen, 1987). The learning objectives largely determine the content and format of instruction, and therefore, these aspects must be coordinated appropriately. Various learning taxonomies can be used to assess instructional alignment, such as Bloom's learning taxonomy (Anderson et al., 2001) or Gagné's learning hierarchies (Gagné, 1968). Primarily, assessing instructional alignment requires a classification of learning goals, learning content and instructional methods.

To this end, ILS&T instruction ideally involves concrete learning goals for language arts and for S&T learning. A widely used classification of the domains that learning goals may focus on is: knowledge, skills and/or attitude (e.g., Bloom and Krathwohl, 1956). Regarding S&T instruction, learning goals can attend to the development of knowledge of the natural and material environment, skills for scientific inquiry and technological design (e.g., defining problems, analysing data), and a critical, curious and investigative attitude (NGSS Lead States, 2013). In language arts instruction, learning goals can address the development of knowledge about language (e.g., text structures and knowledge of reading strategies), language skills (e.g., reading comprehension, the application of reading or writing strategies), and a positive attitude towards language-related learning (Common Core Standards Initiative, 2010). Moreover, language learning activities can serve as a vehicle for S&T learning, for instance, when oral language activities involve talking about S&T phenomena, without offering deliberate support for advancing students' oral language skills. Additionally, the integration of the learning goals can be realized in various ways, depending on the intensity and complexity of subject integration. The framework developed by Gresnigt et al. (2014) proposes a hierarchy of integration approaches: fragmented, connected, fused, multidisciplinary, interdisciplinary and transdisciplinary (see **Table 1**). At the lowest level of integration, the curriculum has separate goals for each subject. At the highest level, the curricular goals transcend the individual disciplines.

Table 1. Classification of integrated language and S&T programs (Gresnigt et al., 2014)

No	Hierarchy of integration
1	Fragmented: Separate and distinct learning goals for the subjects. Often viewed as the traditional way of teaching.
2	Connected: A connection is made between the subjects. The content of the lessons is taught by separate teachers, or the content of the lessons is about one subject.
3	Nested: A skill or knowledge from one subject is targeted within the other subject, but each subject has its own (set of) learning goals and one of the subjects is dominant over the other.
4	Multidisciplinary: Two or more subject areas are organized around the same theme or topic, but the disciplines preserve their identity. Each subject has its own (set of) learning goals, and the subjects are equally important.
5	Interdisciplinary: Skills and concepts are emphasized across the subject areas rather than within the subjects. The (set of) learning goals transcend(s) the individual subjects. The learning goals are (predominantly) taken from subject curricula or schoolbooks and/or are teacher oriented.
6	Transdisciplinary: The curriculum and (set of) learning goals transcend the individual subjects, and the learning goals predominantly include solving real-world problems and/or are student oriented.

Second, what language arts and S&T learning content is taught to students to achieve the learning goals should be considered. Various frameworks have been developed to classify learning content in educational settings. Frameworks from a science inquiry perspective often include declarative or conceptual knowledge (knowing what) and procedures (knowing how, e.g., methods for investigating) (Furtak et al., 2012; McCormick, 1997). A third category that is often distinguished in this context relates to epistemological knowledge (knowing why, the nature of knowledge and how knowledge is created, e.g., knowing how scientists make claims) (Duschl, 2008; Furtak et al., 2012). Learning content can be classified as declarative knowledge (knowing what; e.g., factual or conceptual knowledge), procedures (knowing how, e.g., methods for investigating), or epistemological knowledge (knowing why, the nature of knowledge and how knowledge is created, e.g., knowing how scientists make claims). In many cases, educational interventions attend to more than one learning goal (e.g., knowledge and skills), and more than one type of learning content (e.g., declarative knowledge and procedures).

Third, the instructional method used to achieve the learning goals should be considered. Inquiry- or design-based learning is believed to best suit the nature of science, engineering and technology (Lewis, 2006). Based on a simplified version of the categorization of classroom inquiry developed by Banchi and Bell (2008), a distinction can be made between confirmatory inquiry or design and guided inquiry or design. In confirmatory inquiry or design, students are presented with a question/problem, and follow given procedures to confirm the answer or solution. In guided inquiry or design, students design their own procedures in a self-directed exploration, after being presented with a question or problem. In our view, inquiry- and design-based education should not be confused with unguided discovery learning without any form of direct instruction or transmission of knowledge as directions also have its place in inquiry and design education (Kirschner et al., 2006).

Linguistic activities in S&T instruction, such as argumentation exercises, place high demands on students' cognitive, metacognitive and social abilities (Cheuk, 2016). Therefore, students need (temporary) support from the teacher to carry out tasks that are currently beyond their ability, also referred to as scaffolding (van de Pol et al., 2010). Scaffolding includes a diagnosis of students' needs, which requires teachers to monitor students' learning progress. Next, teachers determine the appropriate follow-up instructional support (i.e., scaffold). This support may involve teachers giving additional explanation, giving modelling examples (e.g., demonstration), or providing cognitive feedback to students. Then, a gradual shift in responsibility from teacher to student takes place as the student becomes more skilled.

Additionally, it can be expected that the duration of the intervention may influence the intended achievements. Students need sufficient instructional time to process information and develop knowledge and skills. Thus, short-term interventions may be less successful in achieving positive learning outcomes, especially when learning goals and content are complex.

Finally, it should be noted that contextual aspects of the student intervention play a role in the implementation, such as individual characteristics of the students (e.g., age, gender, prior knowledge, level of competence in S&T and language arts, see Kyriakides et al., 2018), teacher characteristics (e.g., experience with ILS&T teaching, teachers' beliefs, attitude, and self-efficacy regarding the intervention, see Thurlings et al., 2015) and classroom characteristics (e.g., available time and resources for learning).

Teacher professional development. It is well known that teachers play a pivotal role in the success of educational reform and interventions (Dobber et al., 2017). Scholars have argued that integration of school subjects is a particularly complex undertaking that requires teachers to recognize meaningful connections between the learning processes in both subjects, and therefore requires teacher professional development (TPD; Akerson and Young, 2008; Bradbury, 2014). Literature in the field of content and language integrated learning (CLIL) extensively explored the complexities involved with synthesizing instruction in second language learning and subject areas, amongst which science or STEM (e.g., understanding of effective pedagogical strategies, language proficiency,

Table 2. Overview of study and intervention characteristics that potentially moderate effects

Coded study characteristics			
Research design		Cluster-randomized experiment / quasi-experiment	
Implementation scale (number of students)		Small (< 100) / medium (100-500) / large (> 500)	
Control group		Language-only / S&T-only / separate language and S&T instruction	
Measurement method		Instruments	Independent / researcher-developed
		Time of post-test	Directly after intervention / retention test
Intervention characteristics			
Instructional intervention			
Learning goal	Language [a]	NS / language knowledge / language skills / attitude towards language / language as a means for S&T learning	
	S&T [a]	NS / S&T knowledge, inquiry or design skills / attitude towards S&T	
	Level of integration	NS / fragmented / connected / nested (in S&T or language) / multidisciplinary / interdisciplinary / transdisciplinary	
Learning content	Language [a]	NS / declarative knowledge / procedures / epistemological knowledge	
	S&T [a]	NS / declarative knowledge / procedures / epistemological knowledge	
Instructional method	Learning tasks	NS / confirmatory inquiry/design / guided inquiry/design	
	Monitoring	No / yes	
	Follow-up instructional support [a]	NS / additional explanation / demonstration / cognitive feedback	
	Duration	Number of hours / time span	
TPD intervention			
TPD	Took place in study	Yes / no	
	Duration	Number of hours / time span	
Learning goal [a]		NS, CK (language and/or S&T) / PCK (language and/or S&T) / simple skills, complex skills, attitude	
Learning content [a]		NS / declarative knowledge / procedures / epistemological knowledge	
Instructional method	Standards for intended teacher competencies	None / vague/not clear / clear	
	Modelling of intended instructional practices	Yes / no	
	Opportunities for practice [a]	Isolated from classroom context / situated in classroom context	
		Yes / no	
	Monitoring	Yes / no	
	Follow-up instructional support [a]	NS / additional explanation / demonstration / cognitive feedback	
	Collective participation	NS / individual / team-based	
Attention to contextual conditions	Attention to school leadership and organization (for embedding the innovation)	Yes / no	

[a] A combination of codes is possible.
NS = Not specified.

collaboration, see Kim and Graham, 2022; Pérez Cañado, 2018). However, the existing literature does not yet provide sufficient insight into which aspects of TPD may have an effect on student learning outcomes in the particular context of ILS&T instruction. Below, we draw from the broader literature on effective characteristics of TPD in disciplinary contexts to identify several key features of TPD could potentially affect student learning outcomes in the context of ILS&T instruction. Additionally, in this examination of TPD characteristics, a cognitive psychology perspective is adopted, emphasizing teachers' cognitive processes and decision-making strategies underpinning ILS&T teaching practices.

Because of the complexity of learning to integrate language arts instruction and S&T instruction, it is expected that short-term TPD programs will not suffice. However, the required duration of TPD programs may depend on the prior knowledge and experience of the teachers. Although scholars have often called attention the appropriate duration of TPD programs (e.g., Darling-Hammond et al., 2017; Desimone, 2002), there is no undisputed benchmark.

Much as in the discussion of the important characteristics of the instructional program for students, TPD programs require instructional alignment between the learning goals, learning content and instructional method to be effective (Cohen, 1987). Overall, TPD learning goals can be categorized as attending to the development of knowledge, skills, or attitudes. Two main types of knowledge can be distinguished: content knowledge (CK) and pedagogical content knowledge (PCK) (Fernandez, 2014; Shulman, 1986). In this context, CK refers to domain-specific knowledge of S&T and language arts, while PCK refers to the knowledge of appropriate instructional and pedagogical strategies for teaching subject-matter within the domains. Furthermore, teachers' skills can be categorized as simpler or more complex skills; whereas complex skills require the coordination and integration of knowledge and skills (van Merriënboer and Kirschner, 2017), simple skills do not require the performer to process much information or make many decisions. For example, giving direct instruction to explain an S&T concept can be identified as a relatively simple skill, whereas scaffolding students can be categorized as a complex skill, as it requires teachers to make decisions that vary across different contexts (van Merriënboer and Kirschner, 2017). Finally, learning goals can attend to teachers' attitude(s), for example, towards ILS&T education.

As with the instructional program for students, the learning content of TPD programs can be categorized as declarative knowledge (i.e., factual, conceptual), procedures (i.e., "know-how", methods) or epistemological knowledge (i.e., the nature of knowledge within a domain).

Regarding the instructional method, it is important that there are clear standards for the intended competency development. Such standards can help make the desired outcome of the TPD activities more concrete and assessable, which helps determine whether TPD activities were effective in achieving their goals or not, and to plan for improvement if necessary. Through modelling of the intended instructional practices, best practices can be made explicit to teachers (Borko et al., 2010; Darling-Hammond et al., 2017). Furthermore, researchers have suggested that teachers should engage in practice-based TPD experiences that allow them to immerse themselves in their own learning in order to shift their thinking and teaching practice (Borko et al., 2010; Loucks-Horsley et al., 2009). By offering teachers opportunities for situated practice, the learning tasks are representative of the real task in daily life, which stimulates the transfer of learning (van Merriënboer et al., 2002). As in the instruction for students, teachers can be supported during the TPD activities by providing (a gradual decrease of) scaffolds, by monitoring progress and providing appropriate follow-up instructional support (Fang et al., 2008; van de Pol et al., 2010). In addition, Borko et al. (2010) argued that collaborative practice is an important component of high-quality TPD activities. When teachers can work collaboratively to enhance and reflect on their practice, as well as support and motivate each other, the TPD activities are more effective.

Finally, involving school leadership in TPD activities can enhance the successful implementation of reform efforts (Darling-Hammond et al., 2017; Desimone, 2002). School leaders may actively participate in TPD activities or promote the rationale behind the TPD actively among teachers.

MATERIALS AND METHODS

Selection of Studies

A systematic search was carried out in three databases: Web of Science, Scopus, and ERIC. The search strings included search terms related to

(1) language arts learning (including reading, writing, oral language, and vocabulary),
(2) elementary school and
(3) inquiry- or design-based S&T education.

Due to the complexity of the terminology, two sets of search terms were used, focusing on either inquiry or design learning. To find relevant literature on ILS&T interventions that included inquiry-based learning, the following search string was used: ("5E" OR "inquiry-based learning" OR "inquiry cycle" OR "science inquiry") AND ("language" OR "literacy" OR "vocabulary" OR "reading" OR "writing" OR "oral language") AND ("elementary school" OR "primary school"). The second search string addressed ILS&T interventions that included design-based learning: ("engineering design" OR "technological design") AND ("language" OR "literacy" OR "vocabulary" OR "reading" OR "writing" OR "oral language") AND ("elementary school" OR "primary school"). Several inclusion criteria were identified based on the PICOS framework (population, intervention, comparison, outcome, study type), as can be seen in **Table 3**.

Table 3. PICOS inclusion criteria for study selection

Population	1.	It involves students from kindergarten through 6 (ages 4-12).
	2.	It is undertaken in a school setting to ensure ecological validity.
	3.	It involves a general student population instead of emphasizing student populations with specific characteristics (e.g., second language learners, learning disabilities).
Intervention	4.	It evaluates an intervention that engages students in learning S&T as well as language arts.
	5.	S&T instruction in both conditions involves an inquiry- and/or design-based pedagogy. Educational approaches that are restricted to the transmission of declarative knowledge without paying attention to the practices and nature of S&T are beyond our scope.
Comparison	6.	It examines effect of ILS&T intervention compared to a non-integrated (language and/or S&T) curriculum.
Outcome(s)	7.	It reports quantitative measurements of the effects of the intervention on student learning outcomes for one or more of (a) knowledge, (b) skills, or (c) attitudes in relation to language and/or S&T.
Study type	8.	It describes a two-group pre-test post-test design.
	9.	It was peer-reviewed and published in the last 20 years (2000-2022).

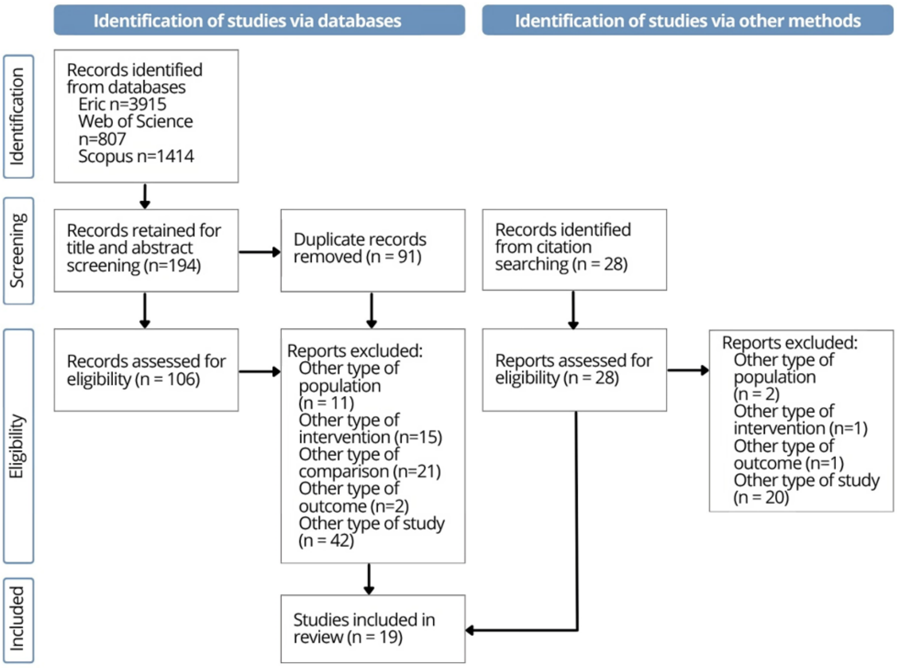
Identification of studies via databases
Identification of studies via other methods
Identification
Screening
Eligibility
Included
Records identified from databases
Eric n=3915
Web of Science n=807
Scopus n=1414
Records retained for title and abstract screening (n=194)
Duplicate records removed (n = 91)
Records identified from citation searching (n = 28)
Records assessed for eligibility (n = 106)
Reports excluded:
Other type of population (n = 11)
Other type of intervention (n=15)
Other type of comparison (n=21)
Other type of outcome (n=2)
Other type of study (n = 42)
Records assessed for eligibility (n = 28)
Reports excluded:
Other type of population (n = 2)
Other type of intervention (n=1)
Other type of outcome (n=1)
Other type of study (n = 20)
Studies included in review (n = 19)

Figure 1. PRISMA flow diagram of study selection process

Additional relevant literature was found through the “snowball method”, by analysing the reference lists of relevant articles to yield further results. **Figure 1** shows the PRISMA flow chart of the selection process for studies included in this review. A total of 19 studies ended up being included.

Analysis

To answer the first research question (What features characterize studies of and interventions for ILS&T?), all intervention characteristics were coded based on coding rules that were developed by the research team. The coding rules covered all variables that emerged from the theoretical framework listed in **Table 2**. Each code included a description and an example. It should be noted that, although the contextual aspects of the student intervention emerged from the theoretical framework as an intervention characteristic, this was not included as a variable in our study because the information in the articles was too limited. All studies were coded separately by the first author based on the coding rules. When studies provided too little information about certain characteristics, these variables were coded as not specified (NS). To determine interrater agreement, a random sample of 20% of the studies was coded by a second independent researcher, who was first trained in the use of the coding rules. Each coder was given a copy of the full articles, coding sheet, and coding instructions. The total percentage of agreement was 88.6%. Due to the large number of categories that was coded, and the fact that the categories consisted of varying numbers of coding options, it was not possible to calculate an overall Cohen’s kappa.

To answer the second research question (Does ILS&T instruction enhance language and S&T learning compared to language and S&T instruction that is not integrated?) the learning outcomes of the interventions were compared using Cohen’s *d* as the common measure for effect sizes (ESs). Cohen’s *d* expresses the standardized mean difference between experimental and control groups. When a study did not report ESs, we primarily used means and standard deviations from the experimental and control group to calculate the difference in post-test means divided by the pooled standard deviation, per the following formula (Cohen, 1998).

$$Cohen's\ d = (M_2 - M_1)/SD_{pooled}$$

where

$$SD_{pooled} = \sqrt{(SD_1^2 + SD_2^2)/2}\,.$$

When the means and standard deviations were not reported, ESs were calculated based on *F*-tests, *t*-tests (Borenstein et al., 2019), or given values of Pearson’s *r* (Ruscio, 2008) or *Z*-score (Rosenthal and DiMatteo, 2001). In the case of missing data, authors were contacted to attempt to retrieve the data. Several studies reported ESs as partial eta squared (η^2_p), which can be defined as the ratio of variance accounted for by an effect and its associated error variance. For comparison, the partial eta squared values were also converted to Cohen’s *d* values (Cohen, 1998), although the partial eta squared results were also given, as they refer to a different type of effect. Some studies reported more than one outcome measure for the same outcome variable. For these studies, the mean ES was calculated for the outcome variable. If the sample sizes were roughly equal, the unweighted mean was computed. If the sample sizes were different, the mean ESs were weighted by the sample sizes. Likewise, there were multiple studies that reported mean scores for different groups (e.g., for grade 3 and for grade 5). These ESs were combined in the same manner, if the groups received the same intervention under the same conditions.

In terms of homogeneity, we assumed high variability of effects across studies. As can be seen in **Table 4**, the studies varied in terms of study design, scale, and instrumentation.

Moreover, studies used different types of control groups, comparing the ILS&T intervention to S&T-only, language-only, or separate S&T and language instruction. In the context of this review, we considered all three control groups to be a type of "business-as-usual" approach. However, they offer a different type of evidence. Therefore, the computation of an overall ES would not be representative and would not allow for the careful consideration of the conditions under which the studies were carried out. To report combined effects of the studies on students' learning achievements, the mean weighted ESs for all outcome variables were computed for studies with a similar study design (i.e., cluster-randomized, quasi-experimental), by weighting the ESs by their inverse variance (Borenstein et al., 2021; Wilson and Lipsey, 2001). The standard error of estimate (*SE*) was calculated for each *ES*, using the formula below.

$$SE = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{\overline{ES}_{sm}}{2(n_1 + n_2)}}.$$

Then, the *SE* was transformed into a weight, using the formula $1/SE^2$. Finally, the weighted mean ES was calculated by dividing the sum of the weighted ESs by the sum of the weightings:

$$\overline{ES} = \frac{\sum(w \times ES)}{\sum w}.$$

To answer the third research question (Which characteristics of ILS&T interventions are associated with the effects on language and S&T learning?), the intervention characteristics were analysed to identify recurring patterns in the relation between those characteristics and the effects on student achievement.

Whether the interventions were characterized by an appropriate match between the learning goals, learning content and instructional method was investigated first. To determine this, four relations were examined, namely,

(1) S&T learning goal and learning content,
(2) language learning goal and learning content,
(3) S&T learning goal and instructional method (i.e., type of learning task), and
(4) S&T learning content and instructional method (i.e., type of learning content).

The alignment of the language learning goal and content with the instructional method was not assessed, because here, the instructional method referred to the type of learning task which was S&T-related (i.e., confirmatory or guided inquiry/design tasks). For each of these relations, a study could receive 1 point, adding up to a maximum total score of 4. Relations that could not be scored due to missing information, for example, when a study did not specify the learning goal, were coded as not specified (NS). This study would also receive a "missing" total score, because otherwise the total score would indicate that one or more of the relations was deemed not appropriate. The first and second relations addressed the match between the learning goal and learning content for S&T and language, which involved a similar set of scoring rules, but they were scored separately. The subjects' learning goals were deemed an appropriate match in case of learning content that included declarative or epistemological knowledge. When learning goals involved skills, the learning content had to address procedures to be an appropriate match. All types of learning content were considered appropriate for learning goals related to attitude. For the third relationship, it was examined whether the S&T learning goals were appropriately aligned to the type of learning task (i.e., confirmatory or guided inquiry/design tasks). Confirmatory inquiry was not considered appropriate for learning goals that included skills in inquiry/design, because students are following given procedures to confirm an answer or solution. Therefore, students likely do not develop inquiry or design skills in such a setting.

Whether the other intervention characteristics distinguished in **Table 2** were associated with the intervention effects was examined next. To do so, similar procedures were followed as for the analysis of the second research question, by calculating the mean weighted ESs of studies when grouped together by similar intervention characteristics, by weighting the ESs by their inverse variance (Borenstein et al., 2021; Wilson and Lipsey, 2001).

Table 4. Overview of study characteristics and effects on student learning outcomes

Reference	DE	SSS	G	Instrument	TT	A	Control Group	Effects		Country
								S&T	Language	
Biyik and Senel (2019)	QE	38	4	RD (S&T knowledge) I (inquiry or design skills)	A, R	Y	S&T only	Knowledge: *d*=**0.66**[a], *d*=**0.47**[a] (retention test) Inquiry or design skills: *d*=**0.57**[a]		Turkey
Cervetti et al. (2012)	CR	467 to 1.027	4	RD	A	Y	S&T only	Knowledge: *d*=**0.54**[a]	Reading: *d*=0.09[a] Writing: *d*=**0.40** Vocabulary: *d*=**0.13**[a]	USA
Chen et al. (2013)	CR	838	4	RD	A	Y (S&T knowledge), N (writing)	S&T only	Knowledge: *d*=**0.25**		USA
Chen et al. (2016)	QE	72	4	RD	A	Y	S&T only	Knowledge: *d*=**0.77**[a] (η^2_p=0.13) Attitude: *d*=**0.51**[a] (η^2_p=0.06)		Taiwan
Girod and Twyman (2009)	QE	53	2	I	A	Y	S&T only	Knowledge: *d*=**1.35**[a] Attitude: *d*=**0.38**[a]		USA
Guthrie et al. (2000)	CR	162	3, 5	I	A	Y	Separate instruction both	-	Reading motivation: *d*=**0.44**[a]	USA
Guthrie et al. (2004)	QE (matched)	243	3, 5	I[c]	A	N	Separate instruction both	-	Reading: *d*=**1.11** Reading motivation: *d*=**1.2**[a]	USA
Hong et al. (2013)	QE	218	5	RD (S&T knowledge), I (S&T attitude)	A	Y	NS	Knowledge: *d*=**0.91**[a] (η^2_p=0.17) Attitude: *d*=**0.29**[a] (η^2_p=0.29)		Taiwan
Kara and Kingir (2021)	QE	107	4	RD	D, A	Y	S&T only	Knowledge: *d*=**1.12**[a]		Taiwan
Lai and Chan (2020)	QE	118	5	RD	A	Y	S&T only	Knowledge: *d*=**0.32**[a] Attitude: *d*=**0.19**[a]		Taiwan
Lutz et al. (2006)	QE	80	4	RD	A	Y	Language only		Reading: *d*=**0.87**[a]	USA
Mercer et al. (2004)	QE (matched)	230	4	I	NS	Y	Separate instruction both	Knowledge: *d*=**0.39**[a]		UK
Romance and Vitale (2001)[b]	QE (matched)	540	4, 5	I	NS	Y	Separate instruction both	Knowledge: *d*=**0.68**[a]	Reading: *d*=**0.22**[a]	USA
van Keulen and Boendermaker (2020)	QE (matched)	141	3-6	I	A	Y	Language only	Attitude: *d*=**0.22**[a]	Reading: *d*=**0.05**[a]	Netherlands
Vitale and Romance (2011)	QE (matched)	513	1, 2	I	NS	Y	Separate instruction both	Knowledge: *d*=**0.16**[a]	Reading: *d*=**0.52**[a]	USA
Vitale and Romance (2012)	QE (matched)	363	1, 2	I	A	N	Separate instruction both	Knowledge: *d*=**0.94**[a]	Reading: *d*=**0.72**[a]	USA
Wigfield and Guthrie (2004)	QE (matched)	350	3	I	A	Y	Language only		Attitude (reading motivation): *d*=**0.18**[a]	USA
Wright and Gotwals (2017)	QE	147	K	RD	A	Y	Separate instruction both	Knowledge: *d*=**1.17**[a]	Vocabulary: *d*=**1.42**[a]	USA
Yang and Wang (2014)	QE	49	4	RD	D, A	Y	S&T only	Knowledge: *d*=**0.63**[a]		Taiwan

Note: Study characteristics that were not specified in studies are indicated by NS. For study type, CR: Cluster-randomized; QE: Quasi-experimental. For instrument, RD: Researcher developed; I: Independent. Grade levels are equivalents in the US educational system. Time of testing is coded as directly after (A), retention (R) or during (D) the intervention. Bolded effect sizes are statistically significant.
DE = Design; SSS = Student sample size; G = Grade; TT = Time of testing; A = Adjusted for pre-test differences.
[a] Cohen's *d* was calculated by the researcher based on the available data about the study.
[b] Only year 4 findings were included from Romance and Vitale (2001), as the other data did not comply with our inclusion criteria.
[b] Only ESs measured by independent measures were included, as both types of instruments were used in this study.

RESULTS

Research Question 1: Characteristics of the Studies and Interventions

Study characteristics

Table 4 shows an overview of the characteristics of all 19 studies included in this review. The studies were mostly conducted in the USA (*n* = 11). The other studies were conducted in Taiwan (*n* = 5), United Kingdom (*n* = 1), Turkey (*n* = 1) and the Netherlands (*n* = 1). They were evenly distributed with respect to their publication dates: with 6 studies being published between 2000-2006, 5 studies between 2007-2013, and 8 studies between

Table 5. Overview of intervention characteristics

Reference	Learning Goal		Learning Content			Method	
	Language	S&T	Integration level	Language	S&T	Learning task	Intervention duration
Biyik and Senel (2019)	Language as S&T learning	Knowledge	N (in S&T)	NS	NS	NS	24 hours/8 weeks
Cervetti et al. (2012)	Skills	Knowledge	M	DK, P	DK, P	GI	30-40 hours/40 weeks
Chen et al. (2013)	Skills	Knowledge	N (in S&T)	P	DK	NS	NS/8 weeks
Chen et al. (2016)	NS	NS	N (in S&T)	NS	DK	CI	24 hours/ 12 weeks
Girod and Twyman (2009)	Knowledge, skills	Knowledge, inquiry or design skills	I	NS	NS	NS	NS/10 weeks
Guthrie et al. (2000)	Skills	Knowledge, inquiry or design skills	M	P	DK, P	GI	NS/36 weeks
Guthrie et al. (2004)	Skills	Knowledge, inquiry or design skills	M	P	DK, P	GI	90 hours/ 12 weeks
Hong et al. (2013)	Skills	Knowledge, inquiry or design skills	I	P	NS	GI	18 hours/12 weeks
Kara and Kingir (2021)	Skills	Knowledge, inquiry or design skills	M	P	P	GI	34 hours/17 weeks
Lai and Chan (2020)	Skills	Knowledge	N (in S&T)	P	DK, P	NS	27 hours/9 weeks
Lutz et al. (2006)	Skills	Knowledge, inquiry or design skills	M	P	DK, P	GI	90-120 hours/12 weeks
Mercer et al. (2004)	Knowledge, skills	NS	N (in language)	DK, EK	DK	GI	12 hours/23 weeks
Romance and Vitale (2001)	Skills	Knowledge	M	P	DK, P	GI	80 hours/40 weeks
van Keulen and Boendermaker (2020)	Skills	Inquiry or design skills	M	P	P	GI	52 hours/26 weeks
Vitale and Romance (2011)	Skills	Knowledge	M	P	DK, P	GI	30 hours/8 weeks
Vitale and Romance (2012)	Skills	Knowledge, inquiry or design skills	M	NS	DK, P	GI	135 hours/36 weeks
Wigfield and Guthrie (2004)	Skills	Knowledge, inquiry or design skills	M	P	DK, P	GI	NS/12 weeks
Wright and Gotwals (2017)[a]	Skills	Knowledge, inquiry or design skills	I	DK, P	DK, P	CI	30 hours/8 weeks
Yang and Wang (2014)	Skills	Knowledge	N (in S&T)	P	DK	CI	12 hours/4 weeks

Note: NS = Not specified. For integration level, N = nested, M = multidisciplinary, I = interdisciplinary. For learning content, DK = declarative knowledge; P = procedures; EK = epistemological knowledge. For learning task, GI = guided inquiry/design, CI = confirmatory inquiry/design.

[a] *Monitoring* was specified to occur only in Wright (2017). Whether *Follow-up instructional support* was provided was not specified in any of the included studies.

2014-2021. Most studies involved students from the middle years of elementary school (Grades 3, 4 and 5) and adopted a quasi-experimental design *(n* = 16*),* out of which 7 studies had composed the experimental and control groups based on matching procedures at the start of the study. In the other studies (*n* = 9), non-random assignment procedures were used. Few studies adopted the strongest research design (i.e., a cluster-randomized controlled trial; *n* = 3) and only one study included a retention test. Ten studies included a medium student sample size (100-500). The other studies adopted either small (<100) or large (>500) student sample sizes (*n* = 5 and *n* = 4*)*, respectively). The experimental conditions were compared to a control condition where students received either only S&T instruction (*n* = 10), separate language and S&T instruction (*n* = 4), or only language instruction *(n* = 4*).* For one study, the type of instruction in the control group was not specified. Most control groups received instruction according to the regular, district-adopted language arts and/or science curriculum (*n* = 16), with a few exceptions where a different intervention was implemented in the control group (e.g., strategy instruction for reading comprehension without S&T integration, Guthrie et al., 2004). The use of independent or researcher-developed instruments to measure student achievement was evenly distributed in the sample.

Intervention characteristics

The interventions could include both instructional intervention (for students) and TPD (for teachers). **Table 5** shows an overview of the characteristics of the instructional intervention in the 19 studies. The ILS&T instructional

Table 6. Overview of TPD characteristics

Reference	Learning goal(s)	Learning content	Opportunities for practice	Modelling of intended instructional practices	M	Duration	ICP
Guthrie et al. (2000)	PCK	NS	NS	NS	NS	40 hours / missing	Individual
Guthrie et al. (2004)	NS	NS	NS	NS	NS	80 hours / missing	Individual
Lutz et al. (2006)	NS	NS	NS	NS	NS	80 hours / 2 weeks	NS
Mercer et al. (2004)	NS	NS	NS	Yes	NS	8 hours / missing	Individual
Romance and Vitale (2001)	CK, PCK (S&T)	Declarative knowledge, procedures	Yes, situated	NS	NS	60 hours / 13 weeks	Team-based
van Keulen and Boendermaker (2020)	PCK (S&T), complex skills	NS	NS	NS	NS	60 hours / 24 weeks	Team-based
Vitale and Romance (2011)	CK, PCK (S&T)	Declarative knowledge, procedures	NS	NS	Yes	32 hours / missing	Individual
Vitale and Romance (2012)	NS	NS	NS	NS	NS	36 hours / 1 week	NS
Wigfield and Guthrie (2004)	NS	NS	NS	NS	NS	Missing / 2 weeks	Individual
Wright and Gotwals (2017)	CK	NS	NS	NS	NS	36 hours / 1 week	Team-based

Note: NS = Not specified. *Standards for intended competencies*, *follow-up instructional support*, and *attention to school leadership* were not specified to occur in any of the included studies. M = Monitoring; ICP = Individual or collective participation.

interventions most often incorporated multiple learning goals that attended to (a combination of) language skills, specifically writing (*n* = 15) and reading (*n* = 12), and S&T knowledge (*n* = 16) and inquiry or design skills (*n* = 10). Regarding the integration level, most studies adopted a multidisciplinary approach (*n* = 10); none of the studies adopted the lowest or highest integration levels (i.e., connected, transdisciplinary). The language-related learning content largely emphasized procedures (e.g., how to write a report; *n* = 14). Writing activities took many forms, for example journal writing, summarizing, and writing evidence-based explanations. The reading activities addressed second-hand investigations through texts, developing content-area reading strategies, and the understanding of text features and structures, among other things. Fewer studies included learning content that addressed oral language (*n* = 6) and vocabulary (*n* = 3). Examples of oral language activities are constructing scientific arguments through evidence-based reasoning and developing discussion skills. The vocabulary activities attended to, for example, academic and scientific vocabulary. The S&T learning content predominantly related to declarative knowledge (e.g., definition of force; *n* = 14). Unfortunately, in many studies the instructional methods used in the intervention were not specified. What did stand out was that most interventions engaged students in guided inquiry or design activities (*n* = 12) but did not specify that the monitoring of students' learning progress was part of instruction, or that a type of follow-up instructional support (e.g., additional explanation, demonstration) was offered to students. The duration of the interventions was widely spread between 12 to 135 hours, with most interventions lasting between 20-50 hours (*n* = 8). The time span over which the intervention took place ranged from 4 to 40 weeks, with most interventions covering relatively short time spans (12 weeks or less; *n* = 12). Four studies did not indicate the number of hours spent on the ILS&T intervention.

Out of the 19 studies, 10 studies incorporated TPD activities to equip teachers for the ILS&T instructional intervention that was implemented in the study. Table 6 shows an overview of the characteristics of the TPD activities that were described in these 10 studies. The mean duration of the TPD activities was 48 hours (*SD* = 23.8) over a mean time span of 8.2 weeks (*SD* = 10.2). In five studies, the learning goals were not specified. In the other studies, the TPD learning goals were related to pedagogical content knowledge (*n* = 4) for S&T instruction, content knowledge (*n* = 3), and complex teacher skills (developing ILS&T lessons). The learning content as operationalized in the TPD activities was not specified in most studies (*n* = 8). The standards for the intended teacher competencies were not specified in any of these studies. Only one study included opportunities for teachers to practice in their own classrooms (situated practice). In one study, specific mention was made that teachers' progress was continually monitored throughout the duration of the TPD activities. In four studies, teachers collectively participated in the TPD activities as a school team. None of the studies reported the involvement of school leaders in the TPD activities.

Research Question 2: Effects of ILS&T Interventions on Students' Learning Achievement

All ESs that were obtained from the 19 included studies are given in Table 4. Out of all obtained ESs, 16 were statistically significant, all of which were in favour of the treatment group. On average, students who received ILS&T instruction demonstrated higher levels of learning achievement than their peers in the control group, with a mean ES of $d = 0.43$. However, it is important to acknowledge the high variability in the ESs, which ranged from $d = 0.05$ to $d = 1.71$. Moreover, the ESs reflected several different outcome variables (e.g., writing, S&T knowledge, reading motivation). It was therefore worthwhile to perform a closer examination of the ESs (per outcome variable). To this end, Table 7 shows that when ESs are grouped by the varying outcome variables, all mean ESs have a positive direction, meaning that the treatment group outperformed the control group. In almost all studies, students receiving the ILS&T intervention outperformed their peers in the control group for learning achievement in language and S&T. Another distinct finding from Table 7 is the relatively high number of studies that measured an intervention effect on reading, S&T knowledge, and attitude towards S&T learning, compared to the other outcome variables. Moreover, none of the studies investigated the effects on students' oral language achievement. Although some interventions did include oral language activities, the effect on students' corresponding language achievement was not measured. In other words, the evidence on the impact of such interventions on oral language achievement was not strong.

Table 7. Mean effect sizes of ILS&T interventions per outcome measure

Outcome variable	Mean ES (Cohen's *d*)	Number of ESs
Vocabulary	0.20	2
Writing	0.40	1
Reading	0.33	7
Reading motivation	0.62	3
S&T knowledge	0.56	14
Inquiry or design skills	0.57	1
Attitude towards S&T learning	0.31	5

Table 8. Comparison of mean effect sizes of ILS&T interventions for study characteristics

Study characteristic		Mean ES (Cohen's *d*)	Number of ESs	Number of studies
Study design	Cluster-randomized	0.25	6	3
	Quasi-experimental	0.61	27	16
	Matched groups	0.78	15	7
	Non-equivalent groups	0.56	12	9
Sample size	Small (< 100)	0.71	8	5
	Medium (100-500)	0.66	19	10
	Large (> 500)	0.26	6	4
Control group type	Separate language and S&T instruction	0.66	12	4
	S&T-only instruction	0.29	15	10
	Language-only instruction	0.21	4	4
	Not clear	0.59	2	1
Type of instrument	Independent	0.57	17	11 [a]
	Researcher-developed	0.32	16	10 [a]

[a] Some studies used both independent and researcher-developed instruments to measure different outcome variables, which is why the sum of the number of studies adds up to more than 19.

Although the findings above provide evidence for the effects of ILS&T instruction on student achievement, we also considered the wide variation in the characteristics of the studies (e.g., study design) and the implications of this for the strength of the evidence provided by the studies. Hence, Table 8 shows the results of categorizing the studies based on their characteristics and determining the mean ESs for the grouped studies. The mean ES for the quasi-experimental studies was more than twice as large as the mean ES for the cluster-randomized studies, with the largest mean ES found for studies that used matching procedures for composing the experimental and control groups. The studies that included a large sample ($N > 500$) yielded a smaller mean ES than studies with small or medium samples. Table 8 shows that studies that compared the ILS&T intervention to separate language and S&T instruction demonstrated the highest mean ES. All studies with a control group receiving separate language and S&T instruction reported statistically significant ESs in favour of the treatment group. These ESs concerned measures of S&T knowledge (5 ESs), reading (4 ESs), reading motivation (2 ESs) and vocabulary (1 ES). Out of the 10 studies that compared ILS&T interventions to S&T-only instruction, 8 ESs were statistically significant, measuring effects related to S&T knowledge (8 ESs), inquiry or design skills (1 ES), attitude towards

S&T learning (1 ESs), writing (1 ES) and vocabulary (1 ES). The mean ES for the studies that compared the ILS&T intervention to language only instruction was considerably lower than the ES for studies with other types of control groups. Among the studies that included a control group receiving language-only instruction, only half of the ESs were statistically significant: one for reading comprehension, and one for reading motivation. Based on these studies, it is difficult to determine whether the integration of S&T with language arts instruction enhances students' learning achievement in language compared to when they are offered language instruction only. Finally, **Table 8** indicates that ESs that were obtained through measures with independent instruments were twice as high as ESs obtained through measures with researcher-developed instruments.

Research Question 3: Relation Between Intervention Characteristics and Intervention Effects

The characteristics of the ILS&T interventions as described in the 19 studies are given in **Table 5**. Overall, it stands out that few studies provide detailed descriptions of the ILS&T intervention, making analysis of the relation between these intervention characteristics and the intervention effects challenging. We hypothesized that a good alignment between the learning goal, learning content and instructional method of the ILS&T interventions would enhance learning and would therefore lead to better results (higher ESs). **Table 9** shows how well the interventions were aligned with respect to their goals, content, and method. Out of the 19 studies, 11 studies provided sufficient information to determine whether there was an appropriate match between the learning goals, content, and instructional method of the intervention. Out of these 11 studies, only 1 study demonstrated a misalignment between these factors (Wright and Gotwals, 2017).

In that study, the learning goals and content included, among others, inquiry or design skills and (S&T) procedures, while instruction followed a confirmatory inquiry approach (i.e., students followed given procedures), which is not the most appropriate method for the development of such skills. Nevertheless, the study reported very high ESs for student achievement, although this may be explained by the fact that only student achievement in S&T knowledge and vocabulary were measured.

Table 9. Alignment between learning goals, content, and instructional method in ILS&T interventions

Reference	Learning goal – Learning content		Learning goal – Instructional method	Learning content – Instructional method	Total score
	Language	S&T			
Biyik and Senel (2019)	Missing	Missing	Missing	Missing	Missing
Cervetti et al. (2012)	1	1	1	1	4
Chen et al. (2013)	1	1	Missing	Missing	Missing
Chen et al. (2016)	Missing	Missing	Missing	Missing	Missing
Girod and Twyman (2009)	Missing	Missing	Missing	Missing	Missing
Guthrie et al. (2000)	1	1	1	1	4
Guthrie et al. (2004)	1	1	1	1	4
Hong et al. (2013)	1	Missing	1	1	Missing
Kara and Kingir (2021)	1	1	1	1	4
Lai and Chan (2020)	1	1	Missing	Missing	Missing
Lutz et al. (2006)	1	1	1	1	4
Mercer et al. (2004)	0	Missing	1	1	Missing
Romance and Vitale (2001)	1	1	1	1	4
van Keulen and Boendermaker (2020)	1	1	1	1	4
Vitale and Romance (2011)	1	1	1	1	4
Vitale and Romance (2012)	Missing	0	1	1	Missing
Wigfield and Guthrie (2004)	1	1	1	1	4
Wright and Gotwals (2017)	1	1	0	0	2
Yang and Wang (2014)	1	1	1	1	4

Table 10 shows the mean ESs for interventions that adopted a similar level of integration. It can be observed that the mean ES increases with higher levels of integration, indicating that when the two subjects are more intertwined, this is accompanied by higher learning gains compared to non-integrated instruction. It was expected that studies with a nested approach, where learning goals for one subject are dominant over those for the other, would yield higher ESs for the dominant subject. However, because all studies with a nested approach (in both S&T and language) only measured S&T learning outcomes, this hypothesis could not be tested. Furthermore, it should be noted that some mean ESs in **Table 10** are only based on one ES and therefore do not provide very strong evidence of a stable pattern.

Table 10. Mean effect sizes for four levels of integration

Level of integration	S&T learning outcomes			Language learning outcomes		
	Mean ES (Cohen's *d*)	Number of ESs	Number of studies	Mean ES (Cohen's *d*)	Number of ESs	Number of studies
Nested in S&T	.36	8	5	-	-	-
Nested in language	.39	1	1	-	-	-
Multidisciplinary	.56	6	6	.31	12	9
Interdisciplinary	.73	5	3	1.42	1	1

Note: The number of studies includes all outcome variables related to the subjects (e.g., for language: reading, writing, vocabulary), which is why the sum of the number of studies adds up to more than 19.

Regarding the duration of the intervention, Table 11 shows that the mean ES obtained from studies with interventions covering a short time span (12 weeks or less) was higher than the mean ES from studies with interventions covering a longer time span (13 weeks or more). Regarding the other intervention characteristics that were examined, only one study specified that during the intervention, the teachers monitored students' learning progress towards the learning goals. None of the studies specified that the teachers implemented follow-up instructional support during the intervention. Therefore, it is difficult to analyse whether this had any impact on the ESs of these interventions.

Table 11. Mean effect sizes for studies with short or long duration of the intervention, and with and without TPD

Studies	Mean ES (Cohen's *d*)	Number of ESs	Number of studies
Studies with short interventions (12 weeks or less)	.59	20	12
Studies with long interventions (13 weeks or more)	.34	13	7
Studies with TPD	.55	14	9
Studies without TPD	.36	19	10

Finally, the last intervention characteristics concerned the TPD activities that were used to prepare teachers for implementation of the instructional intervention. Table 11 shows that studies that included TPD activities yielded higher ESs on average than studies with no TPD activities. The studies provided too little information about the learning goals, learning content and instructional method of the TPD activities to perform any meaningful analysis about the association of these characteristics with the study ES (see Table 6).

DISCUSSION

This review provided an overview of the effects of ILS&T instruction on student learning achievement in language arts and S&T. Unlike previous reviews that evaluated the effects of language arts and science integration, the current review only included studies that focused on S&T instruction with an inquiry- or design-based pedagogy. Furthermore, this review addressed all aspects of language arts learning, including reading, writing, oral language, and vocabulary. Finally, this review described a comprehensive analysis of the features of studies and interventions, and their association with the reported study effects. An important finding is that many studies lacked a detailed description of the study and intervention characteristics, complicating our analysis. This stresses the need for scholars to provide detailed reports of the design and implementation of intervention studies. In the section below, the main findings of the review are discussed along with the limitations and areas for future research.

Characteristics of the Studies and Interventions

The first research question was: What are the characteristics of the studies and interventions in the experimental literature on ILS&T? The analysis resulted in an overview of the distribution of the corpus of studies with respect to their study and intervention characteristics, which showed wide variation. Moreover, many studies lacked specification of the characteristics of the instructional intervention as well as the TPD activities, particularly regarding the instructional methods. Thus, it is unknown whether either these interventions did not include certain TPD characteristics, such as the monitoring of students' learning process or opportunities for teachers to practice, or merely did not explicitly mention doing this in the publications.

Effects of ILS&T Interventions on Students' Learning Achievement

The second research question addressed the effects of ILS&T interventions on student learning. Similar to other review studies in which the impact of integrated instruction in the context of science and language arts learning was investigated (Bradbury, 2014; Graham et al., 2020; Guo et al., 2016; Hwang et al., 2022) we found evidence for the effectiveness of ILS&T instruction. The analysis showed that on average, students who received ILS&T instruction demonstrated higher levels of learning achievement for all reported outcome variables for language and S&T than their peers in a control group. Even though no statistically significant differences between the treatment and control group were found in a few studies, none reported statistically significant effects in favour of the control group. This is an indication that, although they are preliminary due to the limited scope of the available research, there are indications that ILS&T instruction can improve learning achievement in language arts and S&T and does not harm the learning in either subject.

A comparison of the weighted mean ESs of studies with similar study characteristics revealed that ESs tended to be higher for studies that involved a quasi-experimental design and a control group that received separate language and S&T instruction. As expected, studies with a small sample size (< 100) demonstrated higher ESs, although it has been argued that smaller sample sizes tend to overestimate effects (Cheung and Slavin, 2016). The mean ES obtained from independent instruments was higher than the weighted mean ES obtained from researcher-developed instruments. This finding is surprising, as it was expected to be reversed, due to the alignment of researcher-developed instruments with the intervention (Cheung and Slavin, 2016; Wilson and Lipsey, 2001).

Based on the reported outcome variables in the studies that were included in this review, it can be argued that students who received ILS&T instruction outperformed their peers who received separate language and S&T instruction on all measures of S&T knowledge, reading, reading motivation and vocabulary. When compared to S&T-only instruction, evidence was found that ILS&T enhances students' achievement in S&T knowledge, inquiry and design skills, and attitude towards S&T learning. These studies provide evidence that the addition of language arts instruction to S&T instruction enhances S&T learning. Although some of these studies also measured students' language achievement (i.e., vocabulary, writing), ESs found for language achievement are less meaningful in these studies as it can be expected that students who did not receive language instruction during the intervention would score lower on a measure of language achievement. Finally, this review was unable to provide compelling evidence that ILS&T instruction enhances students' learning achievement when compared to language-only instruction, due to the low number of studies that included a control group receiving language-only instruction. Additionally, few to none of the studies measured students' oral language skills, writing skills, and inquiry/design skills, so we could not confirm that ILS&T instruction is effective for enhancing these outcome variables. The absence of such assessment in the studies may be partially due to the complexity of measuring these skills (see Davey et al., 2015; Dockrell and Marshall, 2015).

From a theoretical perspective, there are many potential benefits of ILS&T instruction for enhancing students' language and S&T learning achievement. Still, the approaches to integration that were described in the studies were often relatively rudimentary (e.g., reading or talking about an S&T topic, developing vocabulary). It can be questioned whether the potential of ILS&T is currently being fully harnessed in these interventions. Moreover, it is difficult to assess this based on the reported ESs, because the instruments that are being used to measure student learning achievement are often not aligned with the complexity of the intended (integrated) learning goal of the intervention. The current review only distinguished between independent and researcher-developed instruments, but more features of the measurement procedure are worthwhile to consider.

First, it is important to align assessment with the purpose and content of the intervention, which unfortunately was not always the case in the studies included in this review. For instance, a study that evaluates an intervention that is meant to develop skills (e.g., reading, inquiry/design skills) but only includes assessment of S&T knowledge may produce a high effect size and falsely give the impression that the intervention was highly effective.

Second, the nature of various types of instruments should be considered. For example, testing S&T knowledge or vocabulary can be done in a relatively straightforward and reliable manner, using multiple-choice or open-ended questions that elicit conceptual knowledge. Measuring students' inquiry/design skills or reading comprehension skills can be more complex and requires a different format, such as performance assessment (Shavelson, 1991). Thus, achieving higher ESs for more complex outcome variables may be more challenging.

Third, it should be considered which outcome variable is measured by instruments, and whether this outcome variable aligns with the intended object of assessment. In our analysis, it was difficult to determine the outcome variable that was measured by an instrument in some studies. For example, Chen et al. (2016) designed an instrument in their study that aimed to measure students' written argumentation skills. However, after examining the assessment rubric or scoring rules that were included, we concluded that the instrument only measured students' S&T knowledge and not students' writing skills in a specified genre, and their argumentation skills (which could be considered an S&T skill). Similarly, we concluded that the MAT- and ITBS-tests that were often used in studies only assessed students' scientific knowledge, rather than "critical thinking skills", as was claimed by the authors. This indicates that at times, simplified instruments are used to measure relatively complex skills, which may imply a lack of suitable instruments. This issue has been widely addressed in the context of language arts assessment. For instance, assessment of reading comprehension is inherently complex, as it must be based on indirect symptoms and artifacts of text comprehension (such as disciplinary knowledge), as comprehension itself is a process that cannot be observed directly (Johnston, 1984; Pearson and Hamm, 2005). A common criticism in literature on reading comprehension assessment is whether it is possible to measure the complex interplay of knowledge, strategies and skills required for reading comprehension using a reading test that mainly contains multiple choice comprehension questions (Francis et al., 2005; Pearson and Hamm, 2005). Similarly, effective writing involves a complex interplay of accurate presentation of information, fluency, syntax, and conventions (e.g., see Isaacson, 1984). Thus, more research is needed to explore the availability and suitability of instruments to measure student learning outcomes in the context of ILS&T instruction.

Relation Between Intervention Characteristics and Effects

The third research question was: How do the intervention characteristics relate to the intervention's effects on student learning? Instructional interventions with good alignment between the learning goal(s), learning content and instructional method did not appear to report higher (or lower) ESs than studies with a missing total score; therefore, it remains difficult to conclude whether well-aligned interventions contribute to higher student learning achievement. A further comparison revealed that ESs tended to be higher when the instructional interventions adopted a higher level of integration (i.e., interdisciplinary), which was also found in previous studies (Gresnigt et al., 2014; Loepp, 1999). Additionally, it was found that the mean ESs were higher for interventions that had a short duration (12 weeks or less). It has often been found that short interventions tend to yield higher ESs, but also lead to short-term improvements in student learning. Measuring long-lasting effects for interventions with a long duration (e.g., lasting a school year) can be more challenging.

It was found that interventions that included TPD activities generally yielded higher ESs. Unfortunately, too little information was provided by the studies regarding the TPD characteristics (i.e., learning goals, learning content, instructional method) to perform further analysis. As the focus of the studies was on the instructional intervention, it might not be surprising that the level of detail about the TPD activities was much lower. Moreover, it is worth bearing in mind that the teachers involved in the study may have had sufficient prior knowledge or experience with ILS&T, rendering TPD unnecessary.

Limitations

A few limitations to the present review warrant note. First, this review only included published studies. However, previous reviews have noted that published studies report higher ESs than unpublished ones (Cheung and Slavin, 2016). For this review, it was not deemed feasible to include grey literature. Second, several limitations apply to the ESs' reliability and precision. The ESs reported by the studies vary in their underlying data, as different instruments and statistical methods were used to calculate the intervention outcomes. The preferred method would be to calculate the ES with the complete and original data, rather than deducing ESs from the values mentioned by the studies. However, in this study the researchers made every effort (within reasonable constraints) to ensure the highest possible reliability and precision of the ESs, as described in the Method section. Third, no attention was paid to the treatment fidelity (see Carroll et al., 2007) of the interventions in this review. Ideally, this information would be included, to determine the degree to which the intervention was implemented in classroom practice as intended. Teachers play an important role in this, as well as teacher educators (i.e., was the intervention implemented as planned?). Low treatment fidelity in a study may be a possible explanation for low ESs that are found.

Future Research

This review has shown that more high-quality research is needed to determine whether ILS&T instruction enhances student learning in language and S&T when compared to non-integrated language and S&T instruction, and to understand why certain intervention studies produce more of the desired effects than others. Three main areas of research require attention, namely, specification of the features of the studies and the interventions, expansion of the variables to be investigated, and the systematic analysis of potential moderators of the effects of ILS&T instruction.

Several hypotheses could not be tested in this review due to lack of specification of the characteristics of the intervention in the included studies. This must be rectified in future studies. The authors therefore call upon researchers to be mindful that they thoroughly report the study and intervention design. Researchers should adequately describe the procedures and substantiate design choices when reporting on intervention studies. At the least, this should include the specification of learning goals, learning content and instructional method, and ideally the features of the contextual conditions for learning as well.

More research is needed to expand the relevant variables investigated as an outcome of ILS&T instruction. This review showed that there is disparity in the number of studies reporting effects for the different outcome variables. For example, the number of studies that measured students' reading achievement was much higher than studies reporting on students' writing or oral language skills. To resolve this issue, future studies should include measurements for all student learning outcomes that were addressed in the intervention, which was not the case in many of the studies that were included in this review.

Future research could also further examine the availability or design of suitable instruments for the measurement of relatively complex outcome measures, such as oral language skills, writing skills, and inquiry or design skills. Moreover, it would be worthwhile to design instruments that measure integrated constructs (i.e., argumentative science writing, oral presentations on scientific experiments or technological designs). In this way, the instrument can capture the goals and nature of the ILS&T intervention and of the intended assessment.

Finally, there are other relevant variables that may be worth addressing in future research. For example, this review synthesized studies that compared the experimental conditions (namely ILS&T interventions) to three types of control conditions: separate language and S&T instruction, language-only instruction, and S&T-only instruction. Future research could include a comparison of an ILS&T interventions to all three types of control groups, to give more substantive insight into the value of integration for both subjects.

Finally, more research is needed to systematically analyse the potential moderators of effects of ILS&T instruction. This review focused on the effects of relatively short interventions, where the teachers who implemented the intervention were supported by a team of researchers and experts. In the long term, such support is not always available in practice. The upscaling of programs for system-wide adoption can pose challenges. Moreover, interventions that are successful in controlled settings may not have the same results when used in real-world settings, due to inadequate fidelity among other things. At this stage, other factors can contribute to the success of the implementation, such as active educational leadership (see Timperley et al., 2008). Many reform efforts fail, resulting in teachers returning to traditional teaching methods (Cohen and Mehta, 2017). Making explicit what these factors are could contribute to higher success rates for reform efforts that go beyond controlled interventions.

Longitudinal studies that investigate the effects of ILS&T interventions are required to reveal long-term effects on students' learning achievement in language and S&T. This is especially desirable for examining more complex learning outcomes, such as reading comprehension skills and inquiry or design skills, which are not developed over the course of a few weeks or months. In the current review, weighted mean ESs were calculated for groups of studies with similar characteristics, but due to the variation in the studies and interventions, we were only able to look at one variable at a time. When more studies about ILS&T interventions are available, a thorough meta-analysis could be performed to gain more systematic insight into the potential moderators of the effects of ILS&T instruction. This would also allow for more thorough analysis of the interaction between study and intervention characteristics. Similarly, it would be interesting to look at the cross-over effects of the instructional and TPD interventions. Based on the current review and the data that were available from the included studies, too little is yet known about the combined impact of student and teacher learning in the context of ILS&T instruction.

CONCLUSION

This review provides valuable insights into the impact of ILS&T interventions on students' learning achievement in S&T and language. Although it remains difficult to determine which approach to ILS&T instruction enhances student learning most, this review has taken an important step in providing an overview of the current literature on this topic. This review distinguishes itself from past reviews by focusing on interventions that encompass all domains of language learning (i.e., reading, writing, oral language, and vocabulary), in line with the elementary language arts curriculum. Moreover, this review focused exclusively on interventions that adopt an inquiry- or design-based pedagogy in S&T instruction, which suits the true nature of science, engineering and technology (Lewis, 2006). By assuming this focus, the interventions that were included in this review are closely aligned to classroom practices that are currently being advocated by educational standards (e.g., NGSS Lead States, 2013). The findings of this review also give insight into the areas that are still unknown and require additional research. Moreover, this study provides a substantiated framework for analysing ILS&T interventions and offers new insights into the content and approach of interventions described in the existing literature. These insights can improve the quality of the design of ILS&T interventions. Most importantly, this study showed that ILS&T instruction is, in most cases, effective in enhancing student achievement for language and S&T when compared to non-integrated instruction.

REFERENCES

References marked with an asterisk indicate studies included in the systematic review.

Akerson, V. L. and Young, T. A. (2008). *Interdisciplinary Language Arts and Science Instruction in Elementary Classrooms: Applying research to practice*. Mahwah (NJ): Lawrence Erlbaum Associates.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J. and Wittrock, M. C. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A revision of Bloom's taxonomy of educational objectives*. Harlow: Longman.

Appleton, K. (2007). Elementary science teaching, in S. K. Abell and N. G. Lederman (eds.), *Handbook of Research on Science Education* (pp. 493-535). Mahwah (NJ): Lawrence Erlbaum Associates.

Asma, L., Walma van der Molen, J. and van Aalderen-Smeets, S. (2011). Primary teachers' attitudes towards science and technology, in M. J. de Vries, H. van Keulen, S. Peters and J. Walma van der Molen (eds.), *Professional Development for Primary Teachers in Science and Technology* (pp. 89-106). Rotterdam: Sense Publishers. https://doi.org/10.1007/978-94-6091-713-4_8

Baker, L. (1991). Metacognition, reading, and science education, in C. M. Santa and D. E. Alvermann (eds.), *Science Learning: Processes and applications* (pp. 2-13). Newark (DL): International Reading Association.

Banchi, H. and Bell, R. (2008). The many levels of inquiry. *Science and Children*, 46(2), 26-29.

* Biyik, B. Y. and Senel, A. (2019). Science notebook practice for science lesson: A research on fourth grades. *Cukurova University Faculty of Education Journal*, 48(2), 1367-1399.

Bloom, B. S. and Krathwohl, D. R. (1956). *Taxonomy of the Educational Objectives*. Philadelphia (PA): David McKay.

Borenstein, M., Cooper, H., Hedges, L. V. and Valentine, J. (2019). Effect sizes for continuous data, in H. Cooper, L. V. Hodges and J. C. Valentine (eds.), *The Handbook of Research Synthesis and Meta-Analysis* (pp. 221-235). New York City (NY): SAGE.

Borenstein, M., Hedges, L. V., Higgins, J. P. and Rothstein, H. R. (2021). *Introduction to Meta-Analysis*. Hoboken (NJ): John Wiley & Sons. https://doi.org/10.1002/9781119558378

Borko, H., Jacobs, J. and Koellner, K. (2010). Contemporary approaches to teacher professional development, in P. Peterson, E. Baker and B. McGaw (eds.), *International Encyclopaedia of Education* (pp. 548-556). Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00654-0

Bradbury, L. U. (2014). Linking science and language arts: A review of the literature which compares integrated versus non-integrated approaches. *Journal of Science Teacher Education*, 25(4), 465-488. https://doi.org/10.1007/s10972-013-9368-6

Brophy, J. and Alleman, J. (1991). A caveat: Curriculum integration isn't always a good idea. *Educational Leadership*, 49(2), 66.

Cakir, M. (2008). Constructivist approaches to learning in science and their implications for science pedagogy: A literature review. *International Journal of Environmental & Science Education*, 3(4), 193-206.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J. and Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, 2, 40. https://doi.org/10.1186/1748-5908-2-40

* Cervetti, G. N., Barber, J., Dorph, R., Pearson, D. P. and Goldschmidt, P. G. (2012). The impact of an integrated approach to science and literacy in elementary school classrooms. *Journal of Research in Science Teaching*, 49(5), 631-658. https://doi.org/10.1002/tea.21015

Chen, H.-T., Wang, H.-H., Lin, H.-S., P. Lawrenz, F. and Hong, Z.-R. (2014). Longitudinal study of an after-school, inquiry-based science intervention on low-achieving children's affective perceptions of learning science. *International Journal of Science Education*, 36(13), 2133-2156. https://doi.org/10.1080/09500693.2014.910630

* Chen, H.-T., Wang, H.-H., Lu, Y.-Y., Lin, H.-S. and Hong, Z.-R. (2016). Using a modified argument-driven inquiry to promote elementary school students' engagement in learning science and argumentation. *International Journal of Science Education*, 38(2), 170-191. https://doi.org/10.1080/09500693.2015.1134849

* Chen, Y.-C., Hand, B. and McDowell, L. (2013). The effects of writing-to-learn activities on elementary students' conceptual understanding: Learning about force and motion through writing to older peers. *Science Education*, 97(5), 745-771. https://doi.org/10.1002/sce.21067

Cheuk, T. (2016). Discourse practices in the new standards: The role of argumentation in common core-Eranext generation science standards classrooms for English language learners. *Electronic Journal of Science Education*, 20(3), 92-111.

Cheung, A. C. K. and Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283-292. https://doi.org/10.3102/0013189x16656615

Christie, F. (2017). *Genres and Institutions: Functional Perspectives on Educational Discourse*. London: Continuum. https://doi.org/10.1007/978-3-319-02243-7_2

Cohen, D. K. and Mehta, J. D. (2017). Why reform sometimes succeeds: Understanding the conditions that produce reforms that last. *American Educational Research Journal*, 54(4), 644-690. https://doi.org/10.3102/0002831217700078

Cohen, J. (1998). *Statistical Power Analysis for the Behavioural Sciences*. Mahwah (NJ): Lawrence Erlbaum Associates.

Cohen, S. A. (1987). Instructional alignment: Searching for a magic bullet. *Educational Researcher*, 16(8), 16-20. https://doi.org/10.3102/0013189X016008016

Common Core Standards Initiative. (2010). Common Core State Standards-English Language Arts, *Common Core*. Available at: https://www.thecorestandards.org/ELA-Literacy/

Daniels, H. (2001). *Vygotsky and Pedagogy*. Oxfordshire: Routledge. https://doi.org/10.4324/9780203469576

Darling-Hammond, L., Hyler, M. E. and Gardner, M. (2017). *Effective Teacher Professional Development*. Palo Alto (CA): Learning Policy Institute. https://doi.org/10.54300/122.311

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N. and Wise, L. (2015). *Psychometric Considerations for the Next Generation of Performance Assessment*. Princeton (NJ): Centre for K-12 Assessment & Performance Management, Educational Testing Service.

Desimone, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research*, 72(3), 433-479. https://doi.org/10.3102/00346543072003433

Dobber, M., Zwart, R., Tanis, M. and van Oers, B. (2017). Literature review: The role of the teacher in inquiry-based education. *Educational Research Review*, 22, 194-214. https://doi.org/10.1016/j.edurev.2017.09.002

Dockrell, J. E. and Marshall, C. R. (2015). Measurement issues: Assessing language skills in young children. *Child and Adolescent Mental Health*, 20(2), 116-125. https://doi.org/10.1111/camh.12072

Duschl, R. (2008). Science education in three-partharmony balancing conceptual, epistemic, and social learning goals. *Review of Research in Education*, 32(1), 268-291. https://doi.org/10.3102/0091732x07309371

Fang, Z., Lamme, L., Pringle, R., Patrick, J., Sanders, J., Zmach, C., Charbonnet, S. and Henkel, M. (2008). Integrating reading into middle school science: What we did, found and learned. *International Journal of Science Education*, 30(15), 2067-2089. https://doi.org/10.1080/09500690701644266

Fernandez, C. (2014). Knowledge base for teaching and pedagogical content knowledge (PCK): Some useful models and implications for teachers' training. *Problems of Education in the 21st Century*, 60, 79-100. https://doi.org/10.33225/pec/14.60.79

Francis, D. J., Fletcher, J. M., Catts, H. W. and Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension, in S. G. Paris and S. A. Stahl (eds.), *Children's Reading Comprehension and Assessment* (pp. 387-412). Oxfordshire: Routledge.

Furtak, E. M., Seidel, T., Iverson, H. and Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, 82(3), 300-329. https://doi.org/10.3102/0034654312457206

Gagné, R. M. (1968). Learning hierarchies. *Educational Psychologist*, 6, 1-9. https://doi.org/10.1080/00461526809528968

* Girod, M. and Twyman, T. (2009). Comparing the added value of blended science and literacy curricula to inquiry-based science curricula in two 2nd-grade classrooms. *Journal of Elementary Science Education*, 21(3), 13-32. https://doi.org/10.1007/BF03174720

Graham, S., Kiuhara, S. A. and MacKay, M. (2020). The effects of writing on learning in science, social studies, and mathematics: A meta-analysis. *Review of Educational Research*, 90(2), 179-226. https://doi.org/10.3102/0034654320914744

Gresnigt, R., Taconis, R., van Keulen, H., Gravemeijer, K. and Baartman, L. (2014). Promoting science and technology in primary education: a review of integrated curricula. *Studies in Science Education*, 50(1), 47-84. https://doi.org/10.1080/03057267.2013.877694

Guo, Y., Wang, S., Hall, A. H., Breit-Smith, A. and Busch, J. (2016). The effects of science instruction on young children's vocabulary learning: A research synthesis. *Early Childhood Education Journal*, 44(4), 359-367. https://doi.org/10.1007/s10643-015-0721-6

* Guthrie, J. T., Wigfield, A. and VonSecker, C. (2000). Effects of integrated instruction on motivation and strategy use in reading. *Journal of Educational Psychology*, 92(2), 331. https://doi.org/10.1037//0022-0663.92.2.331

* Guthrie, J. T., Wigfield, A., Barbosa, P., Perencevich, K. C., Taboada, A., Davis, M. H., Scafiddi, N. T. and Tonks, S. (2004). Increasing reading comprehension and engagement through concept-oriented reading instruction. *Journal of Educational Psychology*, 96(3), 403-423. https://doi.org/10.1037/0022-0663.96.3.403

Guthrie, J. T., Wigfield, A., Humenick, N. M., Perencevich, K. C., Taboada, A. and Barbosa, P. (2006). Influences of stimulating tasks on reading motivation and comprehension. *The Journal of Educational Research*, 99(4), 232-246. https://doi.org/10.3200/JOER.99.4.232-246

Hapgood, S. and Palincsar, A. (2006). Where literacy and science intersect. *Educational Leadership*, 64(4), 56-60.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W. and DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, 38(2), 53-67. https://doi.org/10.1111/emip.12253

* Hong, Z.-R., Lin, H.-S., Wang, H.-H., Chen, H.-T. and Yang, K.-K. (2013). Promoting and scaffolding elementary school students' attitudes toward science and argumentation through a science and society intervention. *International Journal of Science Education*, 35(10), 1625-1648. https://doi.org/10.1080/09500693.2012.734935

Hwang, H., Cabell, S. Q. and Joyner, R. E. (2022). Effects of integrated literacy and content-area instruction on vocabulary and comprehension in the elementary years: A meta-analysis. *Scientific Studies of Reading*, 26(3), 223-249. https://doi.org/10.1080/10888438.2021.1954005

International Reading Association and National Council of Teachers of English. (1996). Standards for the English Language Arts International Reading Association and the National Council of Teachers of English, *NCTM*. Available at: https://ncte.org/resources/standards/ncte-ira-standards-for-the-english-language-arts/

Isaacson, S. (1984). Evaluating written expression: Issues of reliability, validity, and instructional utility. *Diagnostique*, 9(2), 96-116.

Jenkins, L. L. (2011). Using citizen science beyond teaching science content: a strategy for making science relevant to students' lives. *Cultural Studies of Science Education*, 6(2), 501-508. https://doi.org/10.1007/s11422-010-9304-4

Johnston, P. H. (1984). Assessment in reading, in D. P. Pearson, R. Barr, M. Kamil and P. Mosenthal (eds.), *Handbook of Reading Research* (pp. 147-182). Harlow: Longman.

* Kara, S. and Kingir, S. (2022). Implementation of the model-based science writing heuristic approach in elementary school science. *International Journal of Science and Mathematics Education*, 20(4), 683-703. https://doi.org/10.1007/s10763-021-10191-0

Khalaf, B. K. and Mohammed Zin, Z. B. (2018). Traditional and inquiry-based learning pedagogy: A systematic critical review. *International Journal of Instruction*, 11(4), 545-564. https://doi.org/10.12973/iji.2018.11434a

Kim, H. and Graham, K. M. (2022). CLIL teachers' needs and professional development: A systematic review. *Latin American Journal of Content and Language Integrated Learning*, 15(1), e1515. https://doi.org/10.5294/laclil.2022.15.1.5

Kintsch, W. (2004). The construction-integration model of text comprehension and its implications for instruction, in R. Ruddell and N. Unrau (eds.), *Theoretical Models and Processes of Reading* (pp. 1270-1328). Newark (DL): International Reading Association. https://doi.org/10.1598/0872075028.46

Kirschner, P. A., Sweller, J. and Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75-86. https://doi.org/10.1207/s15326985ep4102_1

Kyriakides, L., Creemers, B., Charalambous, E., Kyriakides, L., Creemers, B. and Charalambous, E. (2018). The impact of student characteristics on student achievement: A review of the literature, in L. Kyriakides, B. Creemers and E. Charalambous (eds.), *Equity and Quality Dimensions in Educational Effectiveness* (pp. 23-49). New York City (NY): Springer. https://doi.org/10.1007/978-3-319-72066-1_2

* Lai, C. S. and Chan, K. L. (2020). Enhancing science learning through science trade book reading for 5th graders. *Journal of Education in Science, Environment and Health*, 6(1), 1-9. https://doi.org/10.21891/jeseh.669294

Leach, J. and Scott, P. (1995). The demands of learning science concepts: Issues of theory and practice. *School Science Review*, 76, 47-51.

Lee, O., Quinn, H. and Valdés, G. (2013). Science and language for English language learners in relation to next generation science standards and with implications for common core state standards for English language arts and mathematics. *Educational Researcher*, 42(4), 223-233. https://doi.org/10.3102/0013189x13480524

Lemke, J. L. (1990). *Talking Science: Language, learning and values*. New York City (NY): Ablex Publishing Corporation.

Lewis, T. (2006). Design and inquiry: Bases for an accommodation between science and technology education in the curriculum? *Journal of Research in Science Teaching*, 43(3), 255-281. https://doi.org/10.1002/tea.20111

Loepp, F. L. (1999). Models of curriculum integration. *The Journal of Technology Studies*, 25(2), 21-25. https://doi.org/10.21061/jots.v25i2.a.6

Loucks-Horsley, S., Stiles, K. E., Mundry, S., Love, N. and Hewson, P. W. (2009). *Designing Professional Development for Teachers of Science and Mathematics*. Thousand Oaks (CA): Corwin Press. https://doi.org/10.4135/9781452219103

* Lutz, S. L., Guthrie, J. T. and Davis, M. H. (2006). Scaffolding for engagement in elementary school reading instruction. *The Journal of Educational Research*, 100(1), 3-20. https://doi.org/10.3200/JOER.100.1.3-20

Martin, M. O., Mullis, I. V., Foy, P. and Stanco, G. M. (2012). *TIMSS 2011 International Results in Science*. Boston (MA): TIMSS & PIRLS International Study Centre.

McCormick, R. (1997). Conceptual and procedural knowledge. *International Journal of Technology and Design Education*, 7(1), 141-159. https://doi.org/10.1023/A:1008819912213

* Mercer, N., Dawes, L., Wegerif, R. and Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. *British Educational Research Journal*, 30(3), 359-377. https://doi.org/10.1080/01411920410001689689

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington (DC): The National Academies Press. https://doi.org/doi:10.17226/18290

Nixon, D. and Akerson, V. L. (2004). Building bridges: Using science as a tool to teach reading and writing. *Educational Action Research*, 12(2), 197-218. https://doi.org/10.1080/09650790400200245

Osborne, J. (2002). Science without literacy: A ship without a sail? *Cambridge Journal of Education*, 32(2), 203-218. https://doi.org/10.1080/03057640220147559

Pearson, P. D. and Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices—Past, present, and future, in S. G. Paris and A. Stahl (eds.), *Children's Reading Comprehension and Assessment* (pp. 13-69). Mahwah (NJ): Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410612762

Pérez Cañado, M. L. (2018). Innovations and challenges in CLIL teacher training. *Theory Into Practice*, 57(3), 212-221. https://doi.org/10.1080/00405841.2018.1492238

* Romance, N. R. and Vitale, M. R. (2001). Implementing an in-depth expanded science model in elementary schools: Multi-year findings, research issues, and policy implications. *International Journal of Science Education*, 23(4), 373-404. https://doi.org/10.1080/09500690116738

Rosenthal, R. and DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, 52(1), 59-82. https://doi.org/10.1146/annurev.psych.52.1.59

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, 13(1), 19-30. https://doi.org/10.1037/1082-989x.13.1.19

Shavelson, R. J. (1991). Performance assessment in science. *Applied Measurement in Education*, 4(4), 347-362. https://doi.org/10.1207/s15324818ame0404_7

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4-14. https://doi.org/10.3102/0013189X015002004

Skehan, P. and Foster, P. (2001). Cognition and tasks, in P. Robinson (ed.), *Cognition and Second Language Instruction* (pp. 183-205). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524780.009

SLO. (2020). Kerndoelen Primair Onderwijs 2006 SLO, *SLO*. Available at: https://www.slo.nl/@18984/kerndoelen-primair-onderwijs-2006/

Stoller, F. (2008). Content-based instruction, in N. H. Hornberger (ed.), *Encyclopaedia of Language and Education* (pp. 1163-1174). New York City (NY): Springer. https://doi.org/10.1007/978-0-387-30424-3_89

Thurlings, M., Evers, A. T. and Vermeulen, M. (2015). Toward a model of explaining teachers' innovative behavior: A literature review. *Review of Educational Research*, 85(3), 430-471. https://doi.org/10.3102/0034654314557949

Timperley, H., Wilson, A., Barrar, H. and Fung, I. (2008). *Teacher Professional Learning and Development*. Brussels: International Academy of Education.

Traianou, A. (2007). *Understanding Teacher Expertise in Primary Science*. Rotterdam: Sense Publishers. https://doi.org/10.1163/9789087903664

van de Pol, J., Volman, M. and Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review*, 22(3), 271-296. https://doi.org/10.1007/s10648-010-9127-6

* van Keulen, H. and Boendermaker, C. (2020). Contributing to reading comprehension through science and technology education. *Design and Technology Education*, 25(2), 117-142.

van Merriënboer, J. J. G. and Kirschner, P. A. (2017). *Ten Steps to Complex Learning: A Systematic Approach to Four-Component Instructional Design*. Oxfordshire: Routledge. https://doi.org/10.4324/9781315113210

van Merriënboer, J. J. G., Clark, R. and Croock, M. (2002). Blueprints for complex learning: The 4C/ID-model. *Educational Technology Research and Development*, 50, 39-61. https://doi.org/10.1007/BF02504993

* Vitale, M. R. and Romance, N. R. (2011). Adaptation of a knowledge-based instructional intervention to accelerate student learning in science and early literacy in grades 1 and 2. *Journal of Curriculum and Instruction*, 5(2). https://doi.org/10.3776/joci.2011.v5n2p79-93

* Vitale, M. R. and Romance, N. R. (2012). Using in-depth science instruction to accelerate student achievement in science and reading comprehension in grades 1-2. *International Journal of Science and Mathematics Education*, 10(2), 457-472. https://doi.org/10.1007/s10763-011-9326-8

* Wigfield, A. and Guthrie, J. T. (1997). Relations of children's motivation for reading to the amount and breadth or their reading. *Journal of Educational Psychology*, 89(3), 420-432. https://doi.org/10.1037/0022-0663.89.3.420

Wilson, D. B. and Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods*, 6(4), 413-429. https://doi.org/10.1037/1082-989X.6.4.413

Wolf, B. and Harbatkin, E. (2023). Making sense of effect sizes: Systematic differences in intervention effect sizes by outcome measure type. *Journal of Research on Educational Effectiveness*, 16(1), 134-161. https://doi.org/10.1080/19345747.2022.2071364

* Wright, T. S. and Gotwals, A. W. (2017). Supporting kindergartners' science talk in the context of an integrated science and disciplinary literacy curriculum. *Elementary School Journal*, 117(3), 513-537. https://doi.org/10.1086/690273

* Yang, H.-T. and Wang, K.-H. (2014). A teaching model for scaffolding 4th grade students' scientific explanation writing. *Research in Science Education*, 44(4), 531-548. https://doi.org/10.1007/s11165-013-9392-8

Yore, L. D. and Treagust, D. F. (2006). Current Realities and Future Possibilities: Language and science literacy—Empowering research and informing instruction. *International Journal of Science Education*, 28(2-3), 291-314. https://doi.org/10.1080/09500690500336973

Yore, L. D., Bisanz, G. L. and Hand, B. M. (2003). Examining the literacy component of science literacy: 25 years of language arts and science research. *International Journal of Science Education*, 25(6), 689-725. https://doi.org/10.1080/09500690305018